Technical Report on Heart Failure Prediction

Kurt Espinosa
kurtjunshean@gmail.com

1 Introduction

The task is two-fold: 1) create a predictive model to classify patients who have (or will have potential) heart failure or not and 2) to determine the top predictive factors for heart failure.

The given dataset *(processed.cleveland.data)* is of the form X, y where X consists of 13 features with 5 numeric and 8 categorical values and y is categorical-valued target with values $[0, 1, 2, 3, 4]$. Thus, we use a supervised learning approach since we have labelled data.

Some challenges in this task are some features have skewed distributions and the dataset size is relatively small. In this work, we present the classification performance of different models and determine the top predictive factors for heart failure.

2 Data Analysis and Processing

The dataset contains 303 observations and 14 features including the target. Our exploratory data analysis (EDA) shows that some features of the dataset have skewed distributions such as **sex, cp, fbs, oldpeak** as can be seen in the plots generated (see Jupyter notebook). We did not perform outlier detection and removal since some models can handle this kind of distribution. The EDA also shows there are no duplicate rows nor missing cells, which is good.

However, we found some features (i.e., **ca** and **thal**) with values ?, which is not one of their category values. Instead of removing the rows with these values, we decided to replace them with the most frequent category value as it would have high probability to have those category value. Besides, there are only a few of these rows.

More than 50% of the target do not have heart failure and the rest are distributed to all the other values, hence it is skewed. Following previous approaches, we convert the target feature **num** into a binary target $[0, 1]$ where the values $[1, 2, 3, 4]$ map to the presence of heart failure (positive) and 0 maps to the absence of heart failure (negative). The resulting classes is still uneven with less positive (139 or 45%) than negative (164 or 55%) instances but much better distribution than before.

We then split the dataset into train (67%), validation (16.5%) and test (16.5%) sets. Having a validation set prevents us from overfitting our model on the dataset.

3 Experiments and Analyses

3.1 Experiments

In any clinical setting and specifically in this task, we want the model to be more sensitive and not miss alerting any patient with a potential heart failure. Therefore, we use the F1-score as our evaluation metric since it considers both the recall/sensitivity and precision of the model. Besides F1-score gives a clearer picture for dataset with imbalanced classes which is slightly the case in this work. We also plot the ROC curve to show the sensitivity or recall (towards the upper-left corner of the curve) at different thresholds.

We compare eight classifiers: logistic regression, decision tree, random forest, k-neighbors, gradient boosting, XGB classifier, LGBM classifier and CatBoost classifier. Random forest in particular is known to perform well on unbalanced dataset.

We performed 10-fold cross validation on these models and the resulting performance is shown in Figure 1 which shows that the best performing model is random forest with F1-score of 0.8316.

Using grid search, we optimised selected parameters of random forest, namely: *n_estimators, max_features, class_weight, min_samples_split, min_samples_leaf*. We then retrained the random forest using the best parameter values to generate

Figure 1: Model Performance Comparison based on 10-fold cross validation score

Class	P	R	F1
negative (0)	0.88	0.73	0.79
positive (1)	0.69	0.86	0.77
weighted avg	0.80	0.78	0.78

Table 1: Precision, recall and F1-score performance of the tuned model on the test set.

the final model.

We obtained an F1-score (weighted avg) of 0.78 on the test set using the final model as shown in Table 1. Importantly, the model shows a higher recall on the positive class (1) at 0.86 which is what we wanted for this task.

3.2 Analyses

Using the final tuned random forest model, we plot the ROC curve on the test set as shown in Figure 2. We can observe that the curve is closer to the upper-left corner at any threshold which shows the sensitivity of the model.

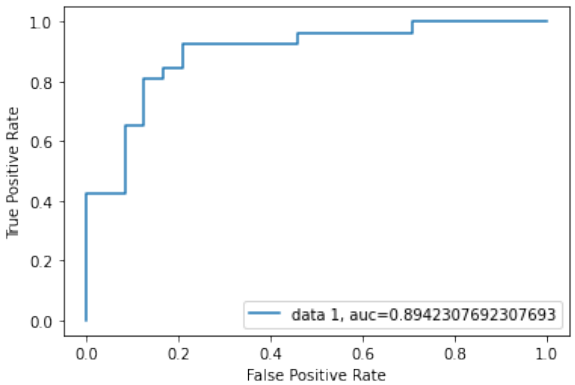

Figure 2: ROC Curve of the tuned model (random forest) on the test set.

Finally, we ranked the features in order of increasing importance as predictors for heart failure as shown in Figure 3. On the one hand, it shows that the top predictor is **thal** followed by **cp** and **thalach**. Based on the documentation, **thal** has 3 category values: 3 for normal, 6 for fixed defect and 5 for reversible defect. Furthermore, **cp** indicates the chest pain type with 4 possible angina-related category values. On the other hand, sex does not seem to be a big factor in predicting heart failure as well as fasting blood sugar (fbs).

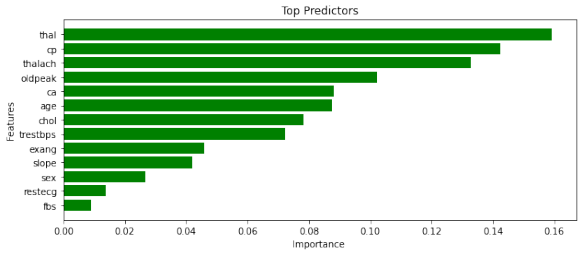

Figure 3: Top predictors for heart failure in decreasing importance.